DOCUMENT RESUME

ED 153 569 EE 009 891

AUTHOR Meskauskas, John A.
TITLE Standard Setting Procedures for Medical Specialty Boards.
PUB DATE Mar 78
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Toronto, Canada, March 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Certification; Higher Education; *Medical Education; *Physicians; Specialization; Speeches; *Standards; *Testing
IDENTIFIERS *Medical Specialty Boards

ABSTRACT

The ways in which standard-setting procedures are carried out in the medical specialty certification area are described. Three experiments that have been conducted with alternatives to currently used normative standards are examined, and reasons why these experiments have had limited success are suggested in this speech. The defensibility of normative as well as absolute standards is also discussed. The predominant pattern in terms of standard-setting for written examination is to define a norming group, typically of candidates from good training backgrounds, and to fail a proportion of that group. Different Boards use different formulas for this, and the exact details of what was chosen are usually not made public. Thus the most widely used approach is normative. When oral examinations are used, the candidate must first pass the (normative) written examination, then the "absolute" oral, which is a hybrid normative-absolute standard. To the extent that they have been challenged in court, the standard- setting procedures used by the boards have been sustained. (SEG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED153569

Standard setting procedures for Medical Specialty Boards

by
John A. Meskauskas
Associate Director for Psychometrics
American Board of Internal Medicine
3930 Chestnut Street
Philadelphia, Pa. 19104

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
John A. Meskauskas
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Prepared for:

Symposium on "Establishing defensible performance expectations in competency-based education programs". American Educational Research Association Annual Meeting, session No. 9.05, March 1978. Toronto, Ontario,

---

The very useful critical comments of Drs. John A. Benson and George W. Webster are gratefully acknowledged.

HE 009871

Standard Setting Procedures for Medical Specialty Boards

This paper has a number of purposes. They are to describe the ways in which standard-setting procedures are carried out in the medical specialty certification arena, to touch briefly on some experiments which have been conducted with alternatives to currently-used normative standards, and to speculate on some of the reasons why these experiments have had limited success. The defensibility of normative as well as absolute standards is discussed.

Background

At the current time, virtually every physician who graduates from a medical school, M.D. diploma in hand, takes some number of years of post-graduate training in one or more medical specialties. The AMA recognizes over 60 specialties, which are based on areas of special knowledge, technology, or practice style differences (e.g., family practice), and offical recognition of competence in the way of a certificate by a Specialty Board is available in over 30. A medical specialist is a physician who by training and experience has competency in a specialty field. This training is obtained within a hospital which has been accredited to provide an approved training experience. The physician studying a specialty is part student, part peer, and part health care provider within a loosely structured (from an educational point of view) hospital setting. Training is only partly didactic, largely consisting of patient care under supervision in a graduated manner so that the trainee acquires more responsibility as he acquires experiences. Such training is generally between 3 and 7 years in length, the extent depending upon the complexity of the knowledge and skills to be acquired and desires of the individual physician.

Specialty Boards have been organized to provide recognition to the public and to colleagues of the special training, experience, and achievement of the specialist practitioners they certify. Specialty Boards are voluntary, non-profit Boards of specialist physicians who set standards of training and develop examination procedures for those who wish to apply to the Board for certification. While the causes for the development of the Specialty Boards are many they are at least partly in response to the fact that any physician who considers himself to be qualified is perfectly free to present himself to the public as a specialist without necessarily undergoing any formal training or any recognized experience in the field. Thus Specialty Boards serve a very important function in attesting to the public the excellence of the preparation of the individuals who they certify.

As part of an increasing concern in the public, government, and also in the profession about the continuing competence of previously-certified practitioners, a number of Specialty Boards have turned to the periodic recertification (at intervals of 6-10 years) of Diplomates to attest to the maintenance of competencies. This is a topic which must be approached carefully, as the public interest and the individual physician's interest could be totally disparate. The Boards tend to be much more careful in the process of denying recertification than they are in not certifying a candidate in the first place.

Examination Procedures

The procedures developed by Specialty Boards for examining their applicants take a number of forms. In every case, an individual's training is checked to make sure that the requisite experiences have been obtained and under favorable auspices. Following verification of that, some Specialty Boards ask

for an assessment by the training director of the clinical competencies of of the trainee. If both of these steps have beer successfully undertaken, the physician is allowed to take an examination or examinations. The certification examinations given by Specialty Boards are one-/or two-day written examinations sometimes supplemented by other examination methodologies such as oral examinations, the reading of slides or graphic materials, case discussions, etc. The examinations are built from a norm-referenced point of view - their purpose is essentially to discriminate between different degrees of achievement in the target population. It has been felt to be impractical to design the written exams from a domain-referenced or criterion-referenced point of view because of the difficulty in defining domains and the perception that a domain-referenced exam, as a summative instrument covering three years' work, would be inordinately long.

The reading of slides and other such "work-sample" methods, while not consciously domain-referenced, are closer to that philosophical point of view than are written examinations. The success of a performance is not dependent on the performance of others but rather on how an individual does with the particular problem(s), and criteria for success are often pre-defined.

Oral examinations based on case material or the examination of live patients fall somewhere between work-sample methods and written examinations because content sampling is sometimes nonsystematic but always limited to very few cases, and the examiner's personal views have a large impact on the outcome. Yet for all that, both orals and work-samples are scored in an "absolute" way in the sense that the outcome is solely dependent on the examinee-content or examinee-content-examiner interaction.

Standard-Setting Processes

The predominant pattern in terms of standard-setting for written examinations is to define a norming group, typically of candidates from good training backgrounds, and to fail a proportion of that group. Different Boards use different formulas for this, and the exact details of what was chosen are usually not made public. Thus the most widely used approach is normative. When oral examinations are used, the candidate must first pass the (normative) written examination, then the "absolute" oral, which is a hybrid normative-absolute standard.

While this writer is unaware of the operational use of absolute standards (referred to later in this paper as content-referenced standards) for written examinations, there is increasing interest in them, particularly for recertification.

Defensibility of the Standard-Setting Process

Several senses of the term "defensibility" need to be distinguished. There is a political sense where one needs to justify one's actions to those who wield political power - a non-issue for specialty boards, because they are not subject to that type of process. Legal defensibility is a concern which increasingly occupies the attention of everyone in our field. To the extent that they have been challenged in court, the standard-setting procedures used by the Boards have been sustained. The recent United States Supreme Court decisions in the Horowitz case and the NEA case have indicated clearly that the actions of responsible professional bodies which act with due process and without bias will not be interfered with by the courts. One can also defend normative standards on the grounds of pragmatics - they most assuredly are easily generated and require a minimum of complexity to implement.

To this writer, the only philosophically defensible standards are those which judge the adequacy of a performance solely on the basis of how well an examinee deals with the content of the examination. (Since judgment is needed to obtain these standards, to call them "absolute" connotes too much-hence we shall use the term "content-referenced" for them.) There is broad interest in content-referenced standards, and some studies have been carried out, but the ultimate outcomes have been disappointing.

Results of Three Experiments with Content-Referenced Standards

On the whole, the area of medical education, as well as allied health practitioners, have been leaders in working toward content-referenced standards. (See, e.g. Reference Note 1, Taylor et.al., (1971), Mattson (1971), Andrew and Hecht (1976). Specialty Boards have also experimented with several methods, but few if any results are published. The experiments have sought to establish content-referenced standards for existing instruments, which are built from a normative point of view.

Two such studies, one by this writer, the other by Dr. Robert Guerin of the National Board of Medical Examiners, used versions of the Nedelsky (1954) method of judging the acceptability of false distractors. The results of these studies were roughly the same, although the work by Dr. Guerin greatly exceeded the work of this writer in procedural adequacy. The judges used were either members of the particular Board or a mixed group of academic and practicing specialists. Seven judges were used in one study, 12 in another. The key findings were that there was a narrow variation of scores on the exams-the standard deviations of the percent-correct scores were in the 5 to 10% range. In contrast, there was a wider range of required minimum levels

of achievement assigned by the various judges. The failure rates of some judges would have failed none while others would have failed up to 58% in one experiment; the range was 0 to 85% in the other experiment. In spite of this, the examination pass/fail cutoff scores set by this method across all judges were reasonable: that is, they were very close to the outcome of the actually-used normative standard. In both cases, the Boards decided not to use the result, for reasons which I will discuss later.

The third experiment was carried out by this writer with clinical simulations called Patient Management Problems. These are latent image feedback simulations which begin with an opening problem statement and then take the individual through a number of sets of actions, providing appropriate feedback to the choices made. The Committee defined the minumum set of actions that were required to pass each case. These actions were both of the positive type, requiring the individual to pick certain choices, as well as the negative type, requiring that the individual avoid certain dangerous, painful or risky options. The specification of these values was done by a group process. The judges read and answered each Patient Management Problem, discussed it among the group, and voted on each option. Approximately two-thirds or higher agreement was obtained, on average, although a simple majority was required. This process resulted in identifying seven to thirteen key decisions out of approximately 50 options per case which must be either correctly chosen or avoided to pass the case. The passing percentages per case varied from 73 to 87% of the examinees. Here again, the outcome was reasonable, but the Board decided not to use this particular type of scoring method, staying with the more accepted scoring methods that had been used in the past.

Comments

In all three experiments, the results closely paralleled the ultimately chosen normative standards, and yet three different groups have independently decided not to use the results. Since many of the members are philosophically committed to "absolute" standards, there is a clear message here. While undoubtedly there is resistance by some to the notion of being able to set standards this way, in this writer's view the basic problems stem from approaches which are insufficiently aware of the complexity of the rater's or judge's task. The result has been that the inability to obtain the kinds of judgments required to utilize the many elegant mathematical models which are available (see, e.g., Millman (1973), Meskauskas (1976) has precluded acceptance of "absolute standards." There are several facets to this problem.

The knowledge base of medicine (some authorities have estimated the half-life of a "bit" of medical knowledge to be 5 years) is constantly changing. Thus, there will be considerable disagreement among the experts as to the proper treatment, diagnosis or other aspects. On a given problem, not all judges will be equally expert, and the individuals who are less expert in a given area will defer to those who strongly hold their point of view or are perceived to be more knowledgeable in a given area. This aspect of the changing state of the art interacting with the human dynamics of a group process introduces a very intractable problem.

The numbers and types of judges required to obtain stable results need to be carefully considered. While wide variation between judges is the rule, it is not clear just what variables account for that. While one could handle this by removing atypical judges, this should be done only when theoretically justifiable and when the distribution of obtained results in normal. (It is

quite possible to get rectangular and even bimodal distributions.) In the absence of a better strategy, one could start with ten judges, calculate the results, and then determine empirically how many more would be necessary to obtain stable results. As to the type of judges, one would think that those with similar characteristics should produce more homogeneous results. In the medical arena, experience suggests variables like specialty training, age, type of clinical problems seen, and area of medical research are likely to important.

The conditions under which the judgments are made are very important. One needs to plan on providing an educational session for the judges, as well as some practice in the judging task, to be able to have any confidence in the stability of the outcome. The question as to whether this should be done in a group session or through the mails is not resolved. While our experience suggests that important differences relating to method of presentation may occur, there is insufficient experimentation to warrant a particular recommendation. The key is to closely monitor the task and to make sure judges understand and are involved in it. If done by the mails, a Delphi-type approach could well be best.

The examination instrument itself may be part of the problem. Examinations designed from a normative point of view may not be amenable to content-referenced standards. Perhaps the distractors necessary to distinguish the incompetent from the barely adequate may not be the same sort as those needed to distinguish between the best answer and all others.

Perhaps the biggest reason for the lack of acceptance of the use of content-referenced standards is what may be termed a revolt against the Procrustean procedures we have used to date. For reasons of mathematical convenience there

is a great deal of reductionism inherent in them, and the outcome is a summation of the bits and pieces. While each of the decisions on single questions or actions does just a little violence to the belief system of the judge, they may be additive rather than zero-sum when aggregated across the entire exam. I think we have been unwrittingly trying to stretch the judges to fit the model. An approach which combines decision theory with psychometrics seems a possibly fruitful avenue for future efforts. Perhaps the work of Paul Hoffman (1960, 1971) and co-workers, which they have termed "policy capturing", could be applied to this problem.

In summary, the experiences to date in this area do not support a hope that one can pluck a model or technique off the shelf and quickly, successfully, apply it in this area. While this realization is a disappointment, it seems likely to lead to better ultimate outcomes. On the positive side, there is a growing support for content-referenced standards and unflagging philosophical support for the importance of the effort. These circumstances, when combined with a well-thought out research effort in the decision processes of standard-setting, will undoubtedly lead to standards which are defensible and supportable by all parties-decision-makers, public, and candidates.

Reference Note

1. Setting standard of competence-The minimum pass level, January 1967. Chicago: University of Illinois, College of Medicine, the Evaluation Unite, Center for the Study of Medical Education.

References

1. Andrew, B.J. and Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. Educ. & Psychol. Meas., 1976, 36, 45-50.

2. Hoffman, P.J. The paramorphic representation of clinical judgment. Psychol. Bull., 1960, 57, 116-131

3. Mattson, D.E. Criterion-related measures in education. J. Med. Educ., 1971, 46, 185-189.

4. Meskauskas, J.A. Evaluation models for cirterion-referenced testing: Views regarding mastery and standard-setting. Rev. Educ. Res., 1976, 46, 133-158.

5. Millman, J. Passing scores and test lengths for domain-referenced measures. Rev. Educ. Res., 1973, 43, 205-216.

6. Nedelsky, L. Absolute grading standards for objective tests. Educ. & Psychol. Meas., 1954, 14, 3-19.

7. Slovic, P., Rorer, L.G., and Hoffman, P.J. Analyzing use of diagnostic signs. Invest. Radiol., 1971, 6, 18-26.

8. Taylor D.D., Reid J.C., Senhauser, D.A., and Shively, J.A. Use of minimum pass levels on pathology examination. J. Med. Educ, 1971, 46, 876-881.